The Polysemy of the Words that Children Learn over Time

Bernardino Casas bcasas@cs.upc.edu [1]

Neus Català ncatala@cs.upc.edu [2]

Ramon Ferrer-i-Cancho rferrericancho@cs.upc.edu [1]

Antoni Hernández-Fernández antonio.hernandez@upc.edu [3]

Jaume Baixeries jbaixer@cs.upc.edu [1]

1. Complexity & Quantitative Linguistics Lab, Departament de Ciències de la Computació, Laboratory for Relational Algorithmics, Complexity and Learning (LARCA), Universitat Politècnica de Catalunya, Barcelona, Catalonia.
2. Complexity & Quantitative Linguistics Lab, Departament de Ciències de la Computació, Center for Language and Speech Technologies and Applications (TALP Research Center), Universitat Politècnica de Catalunya, Barcelona, Catalonia.
3. Complexity & Quantitative Linguistics Lab, Laboratory for Relational Algorithmics, Complexity and Learning (LARCA), Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, Barcelona, Catalonia.

Abstract

Here we study polysemy as a potential learning bias in vocabulary learning in children. We employ a massive set of transcriptions of conversations between children and adults in English, to analyze the evolution of mean polysemy in the words produced by children whose ages range between 10 and 60 months.

Our results show that mean polysemy in children increases over time in two phases, i.e. a fast growth till the 31st month followed by a slower tendency towards adult speech. In contrast, no dependency with time is found in adults. This suggests that children have a preference for non-polysemous words in their early stages of vocabulary acquisition. Our hypothesis is twofold: (a) polysemy is a standalone bias or (b) polysemy is a side-effect of other biases. Interestingly, the bias for low polysemy described above weakens when controlling for syntactic category (noun, verb, adjective or adverb). The pattern of the evolution of polysemy suggests that both hypotheses may apply to some extent, and that (b) would originate from a combination of the well-known preference for nouns and the lower polysemy of nouns with respect to other syntactic categories.

Keywords: child language evolution, vocabulary learning, learning biases, polysemy, quantitative linguistics

Introduction

Children are exposed to millions of words tokens through an accumulation of small interactions grounded in context, immersed in a *sea of words* that they learn early after their birth [31]. During this process, some words are learned first instead of others. Many biases have been hypothesized in the literature in order to explain why some words are learned earlier [32]. For instance, a preference for (a) basic taxonomic level, i.e. less generic words [27, 39], (b) syntactic category, e.g. children learn first nouns and later verbs [12, 13], (c) contextual distinctiveness in space, time or linguistic environment [31], (d) frequent words [15], (e) neighbourhood density, i.e. number of words that sound similar to a given word [37, 38], (f) in-degree of the word in free association norms [19, 18] or (g) mutual information when associating a new word to a meaning [11].

The interest in vocabulary learning biases goes beyond child language research: it has been hypothesized that early stages of language have left traces of simple forms of language [3, 22], child language being one example. On the one hand, word learning biases in children suggest constraints that human language faced at its very origin [32]. On the other hand, the factors by which these learning biases are overridden suggest processes by which communication systems achieve higher linguistic complexity [32].

Children's language development follows a predictable sequence that traditionally has led linguists and psychologists to establish some phases, or stages, in the process of language acquisition [4], with well-known critical periods [23]. At present, we cannot determine exactly the timing of these critical periods for each

individual but many studies suggest the existence of a critical period for syntactic learning approximately between 18 and 36 months of age preceded by a critical period for the phonemic level prior to 12 months [23, 40].

Here we explore the existence of a key and polyhedral bias of language acquisition: word polysemy. Polysemy is the capacity of a word to have multiple meanings (or more than one related sense) and tends to increase as word frequency increases [43, 29, 8, 17]. Recently, the structure of polysemy has been proposed to be a consequence of how children approach the task of building a lexicon with a set of expectations about how words can be used flexibly from early-developing cognitive biases [35].

We analyze the variation in mean polysemy in children over time with the help of a massive database that contains the transcriptions of conversations between children and adults (CHILDES Database). The period of time analyzed (child age) runs from 10 to 60 months. We use two measures of polysemy: the total number of meanings of a word according to the WordNet lexical database, and the number of WordNet meanings of a word that have appeared in an annotated corpus (the SemCor corpus). For further details, see Section Materials and Section Methods.

We will show that mean polysemy in children presents a pattern which is significantly different from that of adults. The mean polysemy in adults remains stable during all the analyzed timespan, whereas that of children starts with a low value, and increases rapidly to almost converge with the mean polysemy of adults.

We consider two possible scenarios to explain these results. The potential preference for non-polysemous words may be:

- **a standalone bias**. This hypothesis would be consistent with the lower uncertainty for those words with respect to their meaning, a factor that may reduce the cognitive cost of learning them as the cost for the listener would be smaller [44].
- **a side-effect** of another bias, for instance, an initial preference for nouns.

Interestingly, we will show that the temporal pattern of polysemy in children weakens when controlling for syntactic category. In order to understand the relationship between polysemy and syntactic category, we have performed two different analyses, that put together, shed light on the hypotheses above. First, we have analyzed the variation in the proportion of four major syntactic categories (noun, verb, adverb and adjective) over the same time period in children and adults. Our results show, again, that children and adults behave differently: initially, children tend to choose a larger proportion of nouns and fewer verbs but by 30 – 33 months these proportions approximate those of adults, which have remained practically constant when speaking to children of different ages. These results agree with previous studies on preference by category (see [32] for a review). Second, we have also analyzed the mean polysemy of verbs and nouns, and the results show that the mean polysemy of nouns is *significantly* lower than that of verbs. These two findings suggest that the tendency of polysemy to increase over time may be mirroring, to some extent, a preference for syntactic category, in particular children's preference for nouns [15]. However, the fact that

the preference for low polysemy words does not disappear completely when controlling for syntactic category, suggests that a standalone bias for low polysemy cannot be discounted.

Results

Evolution of the Polysemy over Time

We estimate the polysemy of a speaker, e.g., a child, from a continuous speech sample, which can be seen as a sequence of N word tokens, $t_1, ..., t_i, ...t_N$. The **mean polysemy** of a sample is defined as the mean of the number of meanings of the i-th token according to its part-of-speech i.e. nouns, adjectives, adverbs and verbs (in these samples, only content words are considered). Samples were obtained from transcripts of recording sessions from the CHILDES database [24] of speech between children and adults (see Section Materials for further details).

From this generic definition, we derive two concrete measures of mean polysemy, depending on how the number of different meanings is calculated for every token:

- **WordNet polysemy**, this is the number of synsets (WordNet meanings) of the word according to the WordNet lexical database.
- **SemCor polysemy**, this is the number of different WordNet synsets associated to a token in the SemCor corpus.

These two measures of polysemy allow one to capture two extremes: the full potential number of synsets of a word (WordNet polysemy) and the actual number of synsets that are used in a corpus (SemCor polysemy), the latter being a more conservative measure of word polysemy motivated by the fact that the number of synsets of a word overestimates, in general, the number of synsets that are known to an average speaker of English or the number of synsets to which a child is exposed.

We analyze the dependency between mean polysemy and time using data from children involved in longitudinal studies. Besides the children role, three adult roles were considered as controls: mothers, fathers and investigators (see Section Methods for further details).

In order to check whether the results are a simple effect of the increase of the production of speech in children as they grow up, we have selected the first n tokens (where n = 50, 75, 100, 125, 150, 175, 250) for each conversation. n is the *length of the conversation*. If a sample does not contain, at least, n tokens, the entire transcript is discarded. Our method of word selection implies that an increase in the length of the conversation tends to reduce the number of conversations that can be included in the analysis. However, a length of 50 tokens yields results that resemble, qualitatively, the results that are obtained with other lengths of conversation. Thus, we take a length of 50 tokens as a canonical length.

For every individual speaker we compute the mean polysemy (according to both WordNet and SemCor polysemies) at each point in time, which is defined as the age of the individual (in months) in that recording session. Therefore, we have a collection of pairs ¡time, mean WordNet polysemy¿ and a collection of

pairs ¡time, mean SemCor polysemy¿ for each individual. We study the evolution of mean polysemy over time from two perspectives:

• **a qualitative analysis**: we average the mean polysemy value for all the individuals that have the same role.

• **a quantitative analysis**: for each role, we count the number of individuals that show significant (positive or negative) or non-significant correlations between mean polysemy and age.

The results of the qualitative analysis of the **WordNet polysemy** are shown in the left-hand side of Figure 1 and those of **SemCor polysemy** are shown in the left-hand side of Figure 2. In both cases, we observe a two-phase (fast-slow) growth of the mean polysemy, delimited by a breakpoint. In adults, there is no clear positive nor negative tendency. The breakpoint for mean WordNet polysemy is located at 31.6 ± 0.1 months, and for mean SemCor polysemy, at 31.4 ± 0.1 months. We note that this breakpoint is computed on the average curve for all children (see Subsection Breakpoint Calculation for details on how the breakpoint is calculated).

The results of the quantitative analysis of the **WordNet polysemy** can be seen in the right-hand side of Figure 1, and the raw data in Table 1 (*All categories* row). These results confirm those of the qualitative analysis: almost half of the children (51.8%) show a significant positive correlation (S_+), whereas most adults (91.1% of the mothers, 85.7% of fathers and 86.7% of investigators) show a non-significant tendency. Notice that that the number of significant positive correlations for children is, significantly high ($\uparrow$), and the number of non-significant correlations is significantly low ($\downarrow$), both according to a binomial test (see Subsection Binomial tests for further details).

The results of the quantitative analysis of the **SemCor polysemy** can be seen in the right-hand side picture of Figure 2, and the raw data in Table 2. In this case, we can see that the trend is the same as in the case of WordNet polysemy. We can also see that a remarkably high number of correlations between mean polysemy and time, are positive and significant in children (51.8%) whereas an overwhelming majority of these correlations are non-significant in all adult categories (85.7% for mothers, 92.9% for fathers and 93.3% for investigators). Finally, we note that the average values for mean polysemy differ depending on the source: an average polysemy of about 6 synsets is found in adults when the source of the polysemy is the SemCor corpus, compared to an approximate average of about 10 synsets when the source is the WordNet database (see Figure 1 and Figure 2).

We have also considered the possibility that these results were a consequence of the dominance of a single morpho-syntactic category over the whole set of tokens. Therefore, we have controlled for category (part-of-speech): we have checked whether the observed pattern also holds for individual categories.

We have taken the same amount of tokens from each recording session but selecting only tokens belonging to one of the four target syntactic categories: nouns, verbs, adjectives and adverbs. Tokens that were not in the target category in their context of use, were discarded. This implied that, if a session did not contain the required number of tokens of the same category, it was discarded.

The results of the qualitative analysis of separate syntactic categories for the

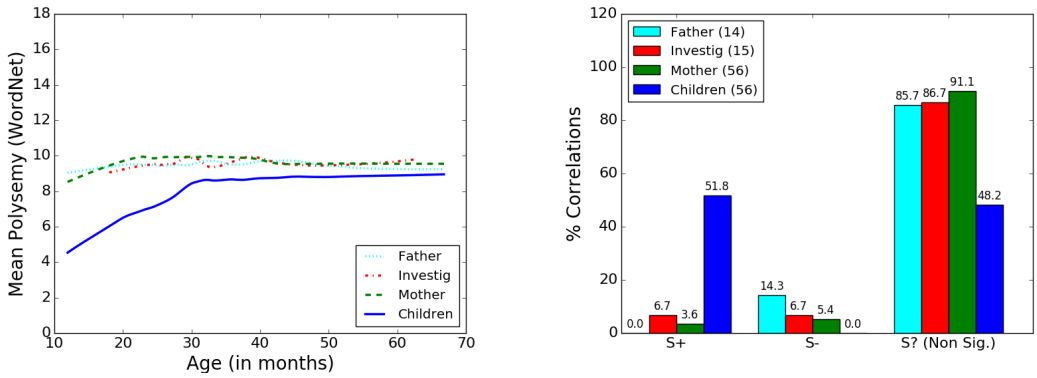

Figure 1. Evolution of the **WordNet polysemy** by children's age in content words. Left: average dependency between mean polysemy and age for each role. The breakpoint in the growth of the mean polysemy of children is located at 31.6 ± 0.1 months. Right: percentage of correlations between mean polysemy and age (S_+: positive significant; S_- negative significant; $S_?$: non-significant). The value above every bar is the percentage of correlations for this role. The number of conversations used to calculate the correlations for every role appears between parentheses in the legend. Length of the conversations: 50 tokens.

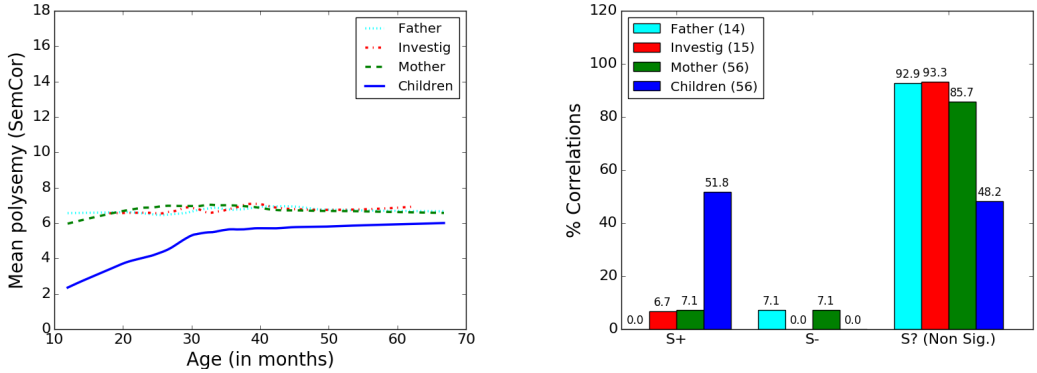

Figure 2. Evolution of the **SemCor polysemy** by children's age in content words The format and length of conversations are the same as in Figure 1. The breakpoint in the growth of mean polysemy of children is located at 31.4 ± 0.1 months.

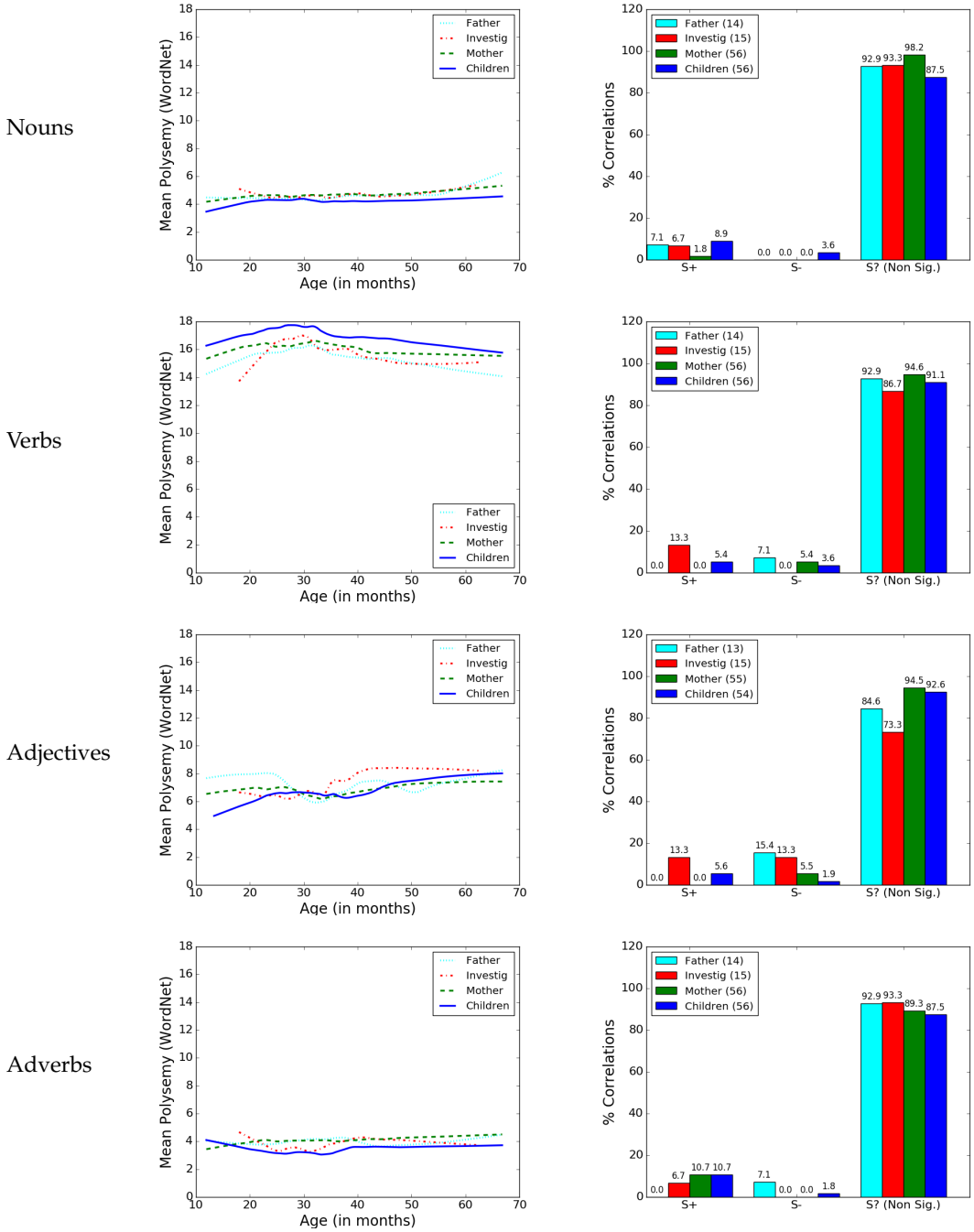

Figure 3. Evolution of the **WordNet polysemy** by age for only nouns, only verbs, only adjectives and only adverbs. Left figures: average dependency between mean polysemy and age of children of each role. Right figures: percentage of correlations between mean polysemy and age (S_+: positive significant; S_- negative significant; $S_?$: non-significant). The value above every bar is the percentage of correlations for this role. The number of conversations used to calculate the correlations for every role appears between parentheses in the legend. Length of the conversations: 50 tokens.

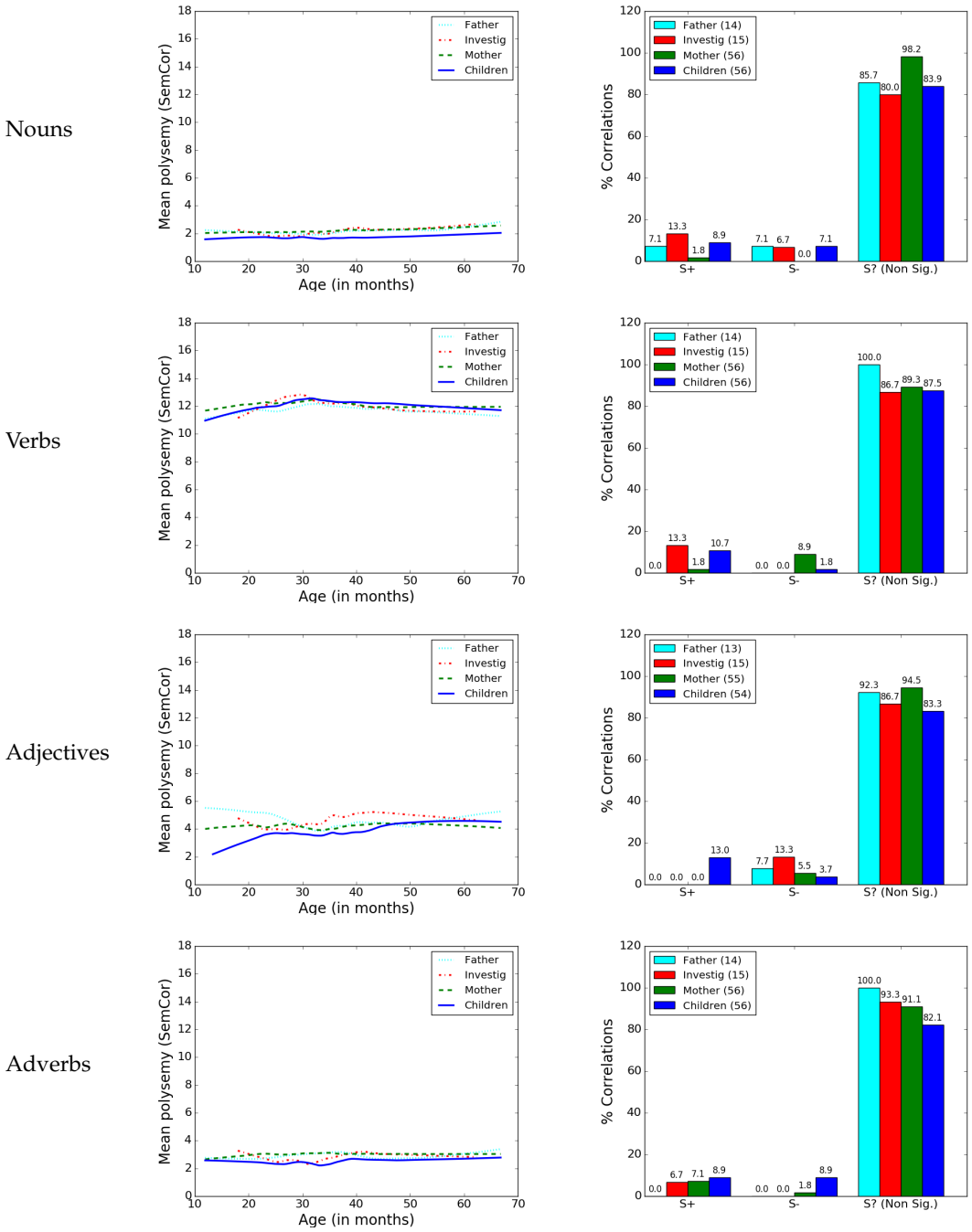

Figure 4. Evolution of the **SemCor polysemy** by age for only nouns, only verbs, only adjectives and only adverbs. The format and length of conversations are the same as in Figure 3.

Category	Role	N	C_+	C_-	S_+	S_-	$S_?$
All categories	Children	56	$\uparrow$ 51	$\downarrow$ 5	$\uparrow$ 27	0	$\downarrow$ 29
	Mother	56	29	27	2	4	50
	Father	15	6	9	0	1	14
	Investigator	15	9	6	1	1	13
Nouns	Children	56	$\uparrow$ 36	$\downarrow$ 20	$\uparrow$ 5	3	$\downarrow$ 48
	Mother	56	34	22	0	0	56
	Father	15	10	5	$\uparrow$ 3	0	$\downarrow$ 12
	Investigator	15	9	6	0	0	15
Verbs	Children	56	25	31	2	4	50
	Mother	56	27	29	2	2	52
	Father	15	6	9	0	1	14
	Investigator	15	5	10	1	0	14
Adjectives	Children	54	$\uparrow$ 37	$\downarrow$ 17	2	2	50
	Mother	55	21	34	0	$\uparrow$ 4	51
	Father	13	6	7	0	0	13
	Investigator	15	6	9	1	2	$\downarrow$ 12
Adverbs	Children	56	26	30	4	3	$\downarrow$ 49
	Mother	56	34	22	$\uparrow$ 6	1	$\downarrow$ 49
	Father	15	8	7	0	1	14
	Investigator	15	8	7	0	0	15

Table 1
*Raw data on the evolution of **WordNet polysemy** over time taking into account the syntactic category and role. N: Number of individuals; C_+: Positive correlations; C_-: Negative correlations; S_+: Positive significant; S_-: Negative significant; $S_?$: Neither positive nor negative significant. Arrows indicate if the counts are significantly high ($\uparrow$) or significantly low ($\downarrow$) according to a binomial test for a given category and role (see the Section Methods for further details about this test). Length of the conversations: 50 tokens.*

mean WordNet polysemy can be found in Figure 3, and the quantitative analyses are shown in Figure 3 and Table 1. Comparing these results with those in Figure 1, the tendency for mean polysemy to increase in children seems to blur, since most children show a non-significant correlation between mean polysemy and time when single syntactic categories are taken alone. Only nouns (8.9%) and adverbs (10.7%) exhibit a relevant number of significant positive correlations (C_+), which are in fact significantly high according to a binomial test. However, these percentages contrast with the percentage of 51.8% when all categories were considered together.

A similar tendency is observed when we analyze the mean SemCor polysemy controlled by category. The qualitative results can be found in Figure 4, and the quantitative results are shown in Figure 4 and Table 2. Again, the tendency for mean polysemy to increase in children is not as clear as when all categories were taken together. However, in this case, it is found that in all categories, the

Category	Role	N	C_+	C_-	S_+	S_-	$S_?$
All categories	Children	56	$\uparrow$ 53	$\downarrow$ 3	$\uparrow$ 29	0	$\downarrow$ 27
	Mother	56	30	26	4	4	$\downarrow$ 48
	Father	14	9	5	0	1	13
	Investigator	15	9	6	1	0	14
Nouns	Children	56	29	27	$\uparrow$ 5	4	$\downarrow$ 47
	Mother	56	31	25	1	0	55
	Father	14	11	3	1	1	12
	Investigator	15	10	5	2	1	$\downarrow$ 12
Verbs	Children	56	35	21	$\uparrow$ 6	1	$\downarrow$ 49
	Mother	56	31	25	1	$\uparrow$ 5	50
	Father	14	7	7	0	0	14
	Investigator	15	8	7	2	0	13
Adjectives	Children	54	$\uparrow$ 35	$\downarrow$ 19	$\uparrow$ 7	2	$\downarrow$ 45
	Mother	55	25	30	0	3	52
	Father	13	5	8	0	1	12
	Investigator	15	7	8	0	2	13
Adverbs	Children	56	28	28	$\uparrow$ 5	$\uparrow$ 5	$\downarrow$ 46
	Mother	56	$\uparrow$ 37	$\downarrow$ 19	4	1	51
	Father	14	8	6	0	0	14
	Investigator	15	5	10	1	0	14

Table 2
*Raw data on the evolution of **SemCor polysemy** over time taking into account the syntactic category and role. The format and length of conversations are the same as in Table 1.*

percentages of positive significant correlations (C_+) in children are significantly high according to a binomial test (8.9% for nouns, 10.7% for verbs, 13% for adjectives 8.9% and for adverbs), but, as in the previous case, they are far from the results in Figure 2 (51.8%).

As for adults, the results in both cases show no relevant tendency, this is, in all cases the non-significant correlations form the majority. This is exactly the same result that was observed when results were not segmented by syntactic category.

Therefore, the trend that we observed in the first part of our analyses dissipates when syntactic categories are considered separately, but does not disappear completely. To try and explain this finding, we have studied the evolution of the use of these categories over time, as well as the relationship between both variables.

Interaction between polysemy and syntactic category

We have analyzed the evolution of the syntactic categories over time using the same methodology as in the previous section. In this case, the variable that is

being analyzed is the percentage of word tokens of a target category over time.

The results of the qualitative analysis (Figure 5) show that the percentage of nouns decreases over time in children: it starts at 80%, drops to 40% and then, stabilizes. Verbs exhibit an opposite tendency in children: they start at 10% and increase their contribution to 40%, and finally they stabilize. In fact, both nouns and verbs seem to stabilize approximately by the same time point. The breakpoint for the percentage of nouns used is located at 30.0 ± 0.1 months and the breakpoint for the percentage of verbs used is located at 33.0 ± 0.1 months (see Subsection Breakpoint Calculation).

These results are consistent with a well-known phenomenon: children tend to learn nouns first and then verbs [32]. As for the remaining categories (adjectives and adverbs), Figure 5 suggests that they remain stable over time. As looks can be deceiving, stronger conclusions must be explored with the help of the quantitative analysis.

According to Figure 6, 42.9% of children show a significant negative correlation between the mean percentage of nouns and time, and 60.7% show a significant positive correlation between the mean percentage of verbs and time, which confirms our observation of the qualitative results. As for adults, they mostly show non-significant correlations in all syntactic categories.

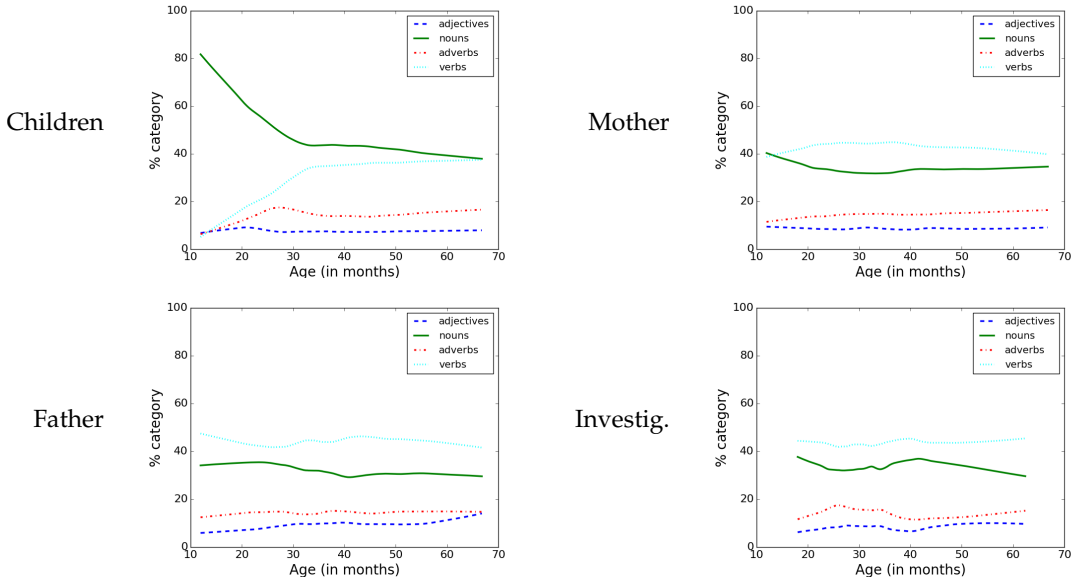

Figure 5. Evolution of the percentage of use of the syntactic categories (adjectives, nouns, adverbs and verbs) by age of children for each role (children, mother, father and investigator). Length of the conversations: 50 tokens. The breakpoint for the percentage of nouns in children is 30.0 ± 0.1 months and for the percentage of verbs in children is 33.0 ± 0.1 months.

Notice that the mean percentage of nouns and verbs stabilize, in children, by an age ranging between 30 and 33 months, that matches the time interval in which children stabilize their mean WordNet and SemCor polysemy (31 months). This suggests that polysemy and syntactic category could be correlated somehow.

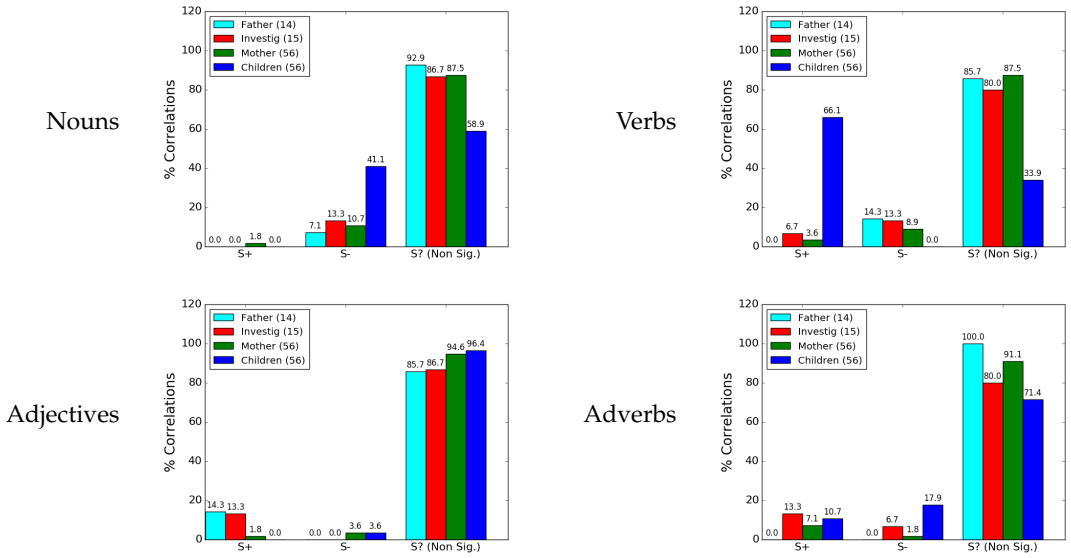

Figure 6. Percentage of correlations (S_+: positive significant; S_- negative significant; $S_?$: non-significant) of each syntactic category by age of children for each role (children, mother, father and investigator). The value above every bar is the percentage of correlations for this role. The number of conversations used to calculate the correlations for every role appears between parentheses in the legend. Length of the conversations: 50 tokens.

The analysis of the abundance of each category suggest that the overall tendency of polysemy to increase could be, at least to some extent, a side-effect of an initial preference for syntactic categories that have low polysemy. For this reason, we calculate the mean polysemy of the tokens of each syntactic category separately, regardless of time. We focus on verbs and nouns, as the other two categories that we have considered (adjectives and adverbs) do not show any relevant tendency, and they represent only a small percentage (less than 20%) of the total of the yielded tokens.

Table 3 shows the mean polysemy of the words (tokens) that have been produced for each syntactic category depending on the role of the speaker. In all roles, verbs have a significantly higher mean polysemy than nouns (a Fisher randomization test gives a p-value $< 10^{-5}$ in all cases; see Subsection Fisher method of randomization for further details about the test).

To sum up, our analysis above shows that children exhibit the following behavior:

1. Approximately between the 30th and 33rd month a breakpoint separates two stages: one of quick maturation followed by another of gradual convergence to adult linguistic behavior. The breakpoints are located at the 31st month for Wordnet and Semcor polysemies, at the 30th month for the percentage of nouns, and at the 33rd month for the percentage of verbs.

2. Up to the breakpoint, their mean WordNet and SemCor polysemy as well as their mean percentage of verbs increase while their mean percentage of nouns

decreases.

3. From the breakpoint onwards, the mean WordNet and SemCor polysemy as well as the mean percentage of verbs and nouns stabilize and tend to converge on those of adults.

Beyond children, our analysis shows that

1. Verbs have a significantly higher mean polysemy than nouns in all roles.

2. With respect to children, adults show a rather stable language production in all analyses over time.

In the next section, we explore the implications of these facts for the origin of the tendency of mean polysemy to increase over time.

Role	POS	Mean	Desv	Max	Tokens	%
Children	Adjective	6.01	5.87	28	90,512	6.66
	Adverb	2.71	2.77	16	211,953	15.61
	Noun	3.65	3.71	33	588,582	43.33
	Verb	15.29	11.45	59	467,172	34.40
Mother	Adjective	6.30	6.11	28	172,685	7.61
	Adverb	3.09	3.39	16	354,798	15.63
	Noun	3.84	3.89	33	742,693	32.72
	Verb	14.48	10.15	59	999,625	44.04
Father	Adjective	6.84	6.62	28	28,299	9.02
	Adverb	3.33	3.47	16	49,187	15.68
	Noun	4.01	4.05	33	101,111	32.24
	Verb	14.15	10.08	59	134,996	43.05
Investigator	Adjective	7.02	6.62	28	13,509	7.41
	Adverb	3.18	3.35	16	29,409	16.12
	Noun	4.15	3.90	33	56,534	30.99
	Verb	13.94	9.91	59	82,950	45.48

Table 3

Statistics about synsets for tokens spoken for each role (Children, Mother, Father and Investigator) of CHILDES. $Mean$*: mean of synsets of tokens of that category spoken for this role.* $Desv$*: standard deviation of the mean.* Max*: maximum number of synsets that have some word for that category.* $Tokens$*: total number of tokens of this category spoken for this role.* $\%$*: percentage of each category over the total words spoken for this role.*

Discussion

In this article we have investigated the polysemy of words as a new potential bias of vocabulary learning. We have shown that the mean polysemy of words increases over time for children, but this effect is not observed in adults. Our finding in children is non-trivial because it is missing in adults interacting with children and thus cannot be attributed to Child Directed Speech [34, 25].

Our finding is also unexpected when considered in the light of principles of language acquisition and facts from quantitative linguistics: the general bias

for high frequency [16] and Zipf's law of meaning distribution, i.e. a positive correlation between frequency and polysemy that concerns both adult and child language [17]. The polysemy effect cannot be explained by a frequency bias in a straightforward fashion since it would imply that children use more polysemous words first, which would contradict our findings presented here. However, our findings could be predicted by the law of meaning distribution and the bias for low frequency that is found across all words but not within specific lexical categories [15].

A critical observation is that the polysemy effect weakens dramatically (but does not disappear completely) when controlling for syntactic category (see Figure 3 and 4). Therefore, we investigated the role that specific categories could have in the polysemy effect. When not focusing on a specific category, the effect could be explained through a combination of three facts:

• Children decrease their proportion of nouns while they increase the proportion of verbs over time (recall Figure 5 and also [32]).

• The mean polysemy of verbs is significantly higher than that of nouns (recall Table 3).

• Nouns and verbs cover the majority of tokens that are produced (recall Figure 5 and Table 3).

Therefore, we suggest two possible explanations, not necessarily mutually exclusive, for the increase in word polysemy over time in children:

1. **Standalone bias**. Children have a preference for less polysemous words, this is, less ambiguous words. This is supported by Zipf's view of polysemy as a cost for the listener [44]. Further support comes from models of Zipf's law that define the listener's effort as the entropy of the meanings of a word, which can be regarded as a distributional measure of polysemy [20].

2. **Side-effect of other biases**. When children learn a language, they begin using more nouns than words from other categories, and, then, they increase the percentage of verbs that they use in their conversations over time. Since the mean polysemy of verbs is *significantly* higher than that of nouns, the mean polysemy increases because the proportion of verbs increases.

On the one hand, the explanation of a side-effect bias is supported by the fact that the positive correlation between mean polysemy and time weakens after controlling for category. In particular, S_+ (the number of speakers with a significant positive correlation between mean polysemy and time) is only significantly high for nouns in a minority of children after controlling for category according to WordNet polysemy (Table 1). On the other hand, the hypothesis of a standalone bias is supported by the different results seen for SemCor polysemy: S_+ reduces substantially after controlling for category but is still significantly high for all syntactic categories according to SemCor polysemy (Table 2). Therefore, both explanations could be valid to a certain extent. In fact, an important question for future research is whether one explanation could be subsumed by the other. Indeed, the standalone bias could provide a more parsimonious explanation: if children prefer less polysemous words in general they will prefer nouns because they are less polysemous. However, we may not be able to reduce all preferences for nouns to polysemy because many nouns are also attractive for

their imageability [26]. Besides, we cannot exclude the possibility that the two explanations above are implications of a different deeper explanation that has escaped us.

Above we have suggested that the bias for low polysemy could weaken within specific categories as a results of it being a side-effect of a preference for nouns. Another reason for the weakening could be a competition with frequency bias, that is stronger within specific categories [15, 18] and is in conflict with the low polysemy bias because the law of meaning distribution predicts that words of low polysemy should have low frequency. Finally, another reason could be that the stand alone bias applies to only certain speakers. This raises the broader question of whether our findings and arguments are valid for all kinds of learners. Interestingly, learners of a second language show an initial tendency for WordNet polysemy to increase over time [7]. This is consistent with our results with L1 learners: as we have shown, children start learning words that are less polysemous, according to two different measures of polysemy (WordNet and SemCor). This suggests that preference for low polysemy is a bias in the vocabulary acquisition process that affects both L1 and L2 learners similarly. An analogous suggestion was made for the bias by which novel words with many similar sounding words are learned more quickly [36].

The breakpoint by the age of 30-33 months in the evolution of polysemy and the percentage of syntactic categories in children coincides in the time-line with the end of the critical period, by the same age, which is traditionally assigned to syntactic development [23, 40, 30], and the production of closed-class words [15]. We believe that the relationship between the stabilization of mean polysemy of children and milestones in the evolution of child language should be the subject of future research.

Conclusions

We have studied the evolution of the polysemy in children, and have put forward polysemy as a learning bias in their vocabulary acquisition. Our main conclusions are:

1. There is a non-trival pattern in the evolution of polysemy over time. Children increase their mean polysemy in two phases: an initial phase with a fast growth of polysemy and a second phase with a slower growth of polysemy. In contrast, adults do not show this tendency.

2. This non trivial pattern weakens when the analysis is segmented by syntactic category.

3. Children show a tendency to learn nouns first and then verbs, which is consistent with previous research [15, 13, 12].

4. Verbs have a significantly higher mean polysemy than nouns in all roles: children and adults.

5. The last two facts could explain the pattern of the evolution of polysemy over time to some extent.

6. That role of a standalone bias for low polysemy cannot be discounted.

7. Our findings and [8] suggest that L1 and L2 learners resemble each other in their dominant biases on polysemy: a preference for low polysemy words prevails in both kinds of learners.

A deeper understanding of why the bias for low polysemy weakens within specific categories and how it interacts with frequency is a challenge for future research. Also, the relationship between the senses that a speaker really knows about a word and its potential number of synsets should be investigated in detail.

Materials

CHILDES database

The longitudinal studies of child language development were taken from the CHILDES database [24]. The majority of corpora within this database are transcripts of conversational interactions among children and adults which occur at a given point in time, which are referred to as **recording sessions** throughout this paper.

The longitudinal studies of child language development from the CHILDES database [24] that were analyzed are the same as in [2] for English language: 60 target children. In this article we only analyze content words (nouns, verbs, adjectives and adverbs).

All the corpora of the CHILDES database are freely available at `http://childes.psy.cmu.edu/data` (accessed 17 December 2012).

Lexical database WordNet

The polysemy of a word in English was obtained by querying the lexical database WordNet [28, 10].

WordNet can be seen as a set of synsets and relationships among them. A synset is the representation of an abstract meaning and is defined as a set of words having (at least) the meaning that the synset stands for. As an example in WordNet, the word *book* is related (among others) with the synset that represents *a written work or composition that has been published*, with the synset that represents *a written version of a play or other dramatic composition; used in preparing for a performance* or the synset that represents *to arrange for and reserve (something for someone else) in advance*. Each pair word-synset contains, as well, the information that corresponds to a syntactic category related to those two elements. For instance, the pair *book* and the synset *a written work or composition that has been published* are related to the category *noun*, whereas the pair *book* and synset *to arrange for and reserve (something for someone else) in advance* are related to the category *verb*.

WordNet has 155,287 lemmas and 117,659 synsets (see `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`), and contains only four main syntactic categories: nouns, verbs, adjectives and adverbs. Words with other syntactic categories are not present in this database (for instance, the article *the* or the preposition *for*).

WordNet is freely available for download at http://wordnet.princeton.edu/wordnet/download/. We use WordNet version 3.0, that it is included in the NLTK platform (Natural Language Toolkit) version 2.0 freely available at http://www.nltk.org.

SemCor corpus

The Semantic Concordance Package (SemCor) is a corpus composed of 352 texts which are a subset of the English Brown Corpus. All words in the corpus were syntactically tagged using Brill's part of speech tagger. The semantical tagging was done manually, mapping all content words to their corresponding synsets in WordNet.

SemCor contains $676,546$ tokens, $234,136$ of which are sense-tagged. This yields $23,341$ different tagged lemmas that represent only content words.

We processed SemCor to count the number of differents WordNet meanings that every pair ¡lemma, syntactic category¿ has in this corpus. All this information was saved to be used in the processing of the recording sessions.

Table 4 shows the percentage of CHILDES lemmas that are present in SemCor corpus. The coverage of SemCor (proportion of SemCor lemmas that appear in WordNet) for verbs, adjectives and adverbs is above 75%, for nouns it is lower, between 53,61% and 59,32%.

SemCor is freely available for download at http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor.

Methods

TreeTagger

TreeTagger [33] was used to determine the syntactic category and the lemma of word tokens from transcripts of speech. Essentially, this tool annotates text with part-of-speech (POS) and lemma information (canonical form). That tool was chosen because it supports many languages, English among others, which could facilitate the extension of our studies to other languages as well in the future.

As an example of the performance of Treetagger, let us consider the word *book* in the sentence *He gave the book to his sister*. TreeTagger tags *book* in this sentence as *noun* and yields, as well, a lemma or canonical form for that word (in this case, the lemma of *book* would be *book*, and the lemma of *gave* would be *give*).

In English, the POS tags that refer to the main categories are:

- Adjectives: JJ, JJR and JJS.
- Adverbs: RB, RBR and RBS.
- Nouns: NN and NNS.
- Verbs: MD, VBD, VBG, VBN, VBP, VBZ, VD, VDD, VDG, VDN, VDP, VDZ, VH, VHD, VHG, VHN, VHP, VHZ, VV, VVD, VVG, VVN, VVP and VVZ .

Role	POS	# typ	# typ sem	% typ	# tok	# tok sem	% tok
Children	n	5,973	2,616	43.80	588,582	315,518	53.61
	v	1,584	1,109	70.01	467,172	443,555	94.94
	a	1,261	731	58.05	90,512	76,102	84.08
	r	369	246	66.66	211,953	159,094	75.06
Mother	n	8,023	3,571	44.51	742,693	439,839	59.22
	v	2,438	1,738	71.29	999,625	965,200	96.56
	a	2,131	1,251	58.70	172,685	155,212	89.88
	r	633	436	68.88	354,798	311,011	87.66
Father	n	4,172	2,279	54.63	101,111	58,040	57.40
	v	1,247	992	79.55	134,996	130,698	96.82
	a	1,013	671	66.24	28,299	26,202	92.59
	r	376	279	74.20	49,187	42,039	85.47
Investigat.	n	2,361	1,490	63.11	56,534	33,538	59.32
	v	785	658	83.82	82,950	79,974	96.41
	a	577	439	76.08	13,509	12,753	94.40
	r	254	206	81.10	29,409	24,868	84.56

Table 4
Statistics of processed lemmas of CHILDES that appear in SemCor by syntactic category according to role (Children, Mother, Father and Investigator). POS: Part-of-speech (n: nouns, v: verbs, a: adjectives, r: adverbs). $\#typ$: number of types. $\#typ\ sem$: number of types that appear in SemCor. $\%typ$: percentage of types that appear in SemCor. $\#tok$: number of tokens. $\#tok\ sem$: number of tokens that appear in SemCor. $\%tok$: percentage of tokens that appear in SemCor.

Processing data

For each recording session, we processed the list with all tokens of the transcript in the same order that have been produced by a certain speaker. For each of these tokens, we proceeded as follows:

1. *Tagging*. Each token in this list was assigned a morpho-syntactic category tag and a lemma by TreeTagger. Therefore, this list contained now triples ¡token, syntactic category, lemma¿.

2. *Discard unprocessable tokens*. We discarded from the previous list those triples whose tokens have non-ascii chars or are CHILDES special tags (”@”, ”xxx”, ”xx”, ”yyy”, ”yy”, ”www”).

3. *Proper noun recognition*. We recognized the proper nouns from a list of proper nouns. Any triple whose token appeared in this list was tagged as proper noun. This list was compiled with information extracted from different sources and can be downloaded at `http://tinyurl.com/polysemy-ne-txt`.

4. *Category filtering*. From the resulting list in the previous step, we only selected those triples whose morpho-syntactic category corresponds to a content word.

5. *Polysemy calculation*. We used two methods to calculate the polysemy for

every ¡token, syntactic category, lemma¿ of the list computed in the previous step:

- *WordNet query*. WordNet was queried to obtain the number of synsets related to each triple ¡token, syntactic category, lemma¿ from the previous list. We first queried WordNet the pair ¡token, syntactic category¿. If this pair was not found, we queried again WordNet the pair ¡lemma, syntactic category¿. If this pair was not found, this token was discarded and not considered for calculations of the WordNet polysemy.
- *SemCor query*. The file with the processed information of SemCor was queried to obtain the number of synsets related to each triple ¡token, syntactic category, lemma¿ from the previous list. We queried the pair ¡lemma, syntactic category¿. If this pair was not found, this token was discarded and not considered for calculations of the SemCor polysemy.

For instance, in the former example where the token *book* was tagged to be a noun, we would only select the 11 synsets yielded by WordNet for the *noun* category. If there were no synsets related to ¡*book*, noun¿ this token would be discarded.

We also performed some extra controls:

1. A **control by conversation length**: from the remaining lists that have passed the previous filters, we selected the first n tokens (where $n = 50, 75, 100, 125, 150, 175, 250$) for each session. If the sample did not have at least n tokens then the entire transcript was discarded.

2. A **control by morpho-syntactic category**: from the remaining list, we only considered those tokens that belong to a specific morpho-syntactic category: noun, adjective, verb or adverb. In the study of the specific morpho-syntactic category we also controlled for length of conversations: we took the first n tokens after applying the morpho-syntactic category filter (tokens that do not belong to the target syntactic category are discarded). As before, if the sample did not have, at least, n tokens, then, the entire transcript was discarded.

Finally, notice that all the processing steps above imply that the analysis that is not restricted to a specific morpho-syntactic category defines the polysemy of a token according to the morpho-syntactic category of the triple. It does not define the polysemy aggregating the polysemy for all content categories somehow (e.g., by summing the polysemy of the token for all content categories). The reason for our choice is to run the raw analysis in as similar a way as possible to that carried out when controlling for morphosyntactic category.

Mathematical Computations

The association between mean polysemy and age was measured with a Spearman rank correlation, $\rho_S(X, Y)$, which is a measure of monotonic dependency between a pair of variables [5]. Although the traditional Pearson correlation has been used to investigate the relationship between polysemy and time in L2 learners [8], Spearman rank correlation has the advantage of being able to capture non-linear dependencies; the traditional Pearson correlation is simply a

measure of linear dependency [14, 9].

To determine whether a correlation was significant or not, a two-sided correlation test with a significance level of $\alpha = 0.05$ was used (a correlation is significant if the p-value does not exceed α).

The smoothing of the plots in Figures 1, 3 and 5 was performed with the non parametric smoother method `lowess` implementated in Python in the library *statsmodels* (see `http://statsmodels.sourceforge.net/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html`) with parameters $frac = 1./3, it = 0$.

All participants with less than m^* time points were excluded from the analyses. m^* is the minimum number of points that are needed for significance by a two-sided correlation test between two vectors X and Y. m^* is the smallest value of m satisfying the condition [21]

$$2/(m!) \leq \alpha, \tag{1}$$

where α is the significance level. With $\alpha = 0.05$ then $m^* = 5$.

Fisher method of randomization

To assess whether verbs have a higher mean polysemy than another category for a given role, we performed a Fisher randomization test [5]. The statistic used is the absolute value of the difference between the mean polysemy of the verb tokens and the mean polysemy of the tokens of the other category. The test checks whether the absolute difference is significantly high. The p-value of the test was determined by means of a Monte Carlo procedure over 10^5 randomizations.

Binomial tests

The $\uparrow$ and $\downarrow$ arrows in Tables 1 and 2 indicate if a certain number is significantly high or significantly low. These arrows were determined with the help of binomial tests [6, 2]. Suppose that N is the number of individuals with at least m^* points of time and α is the significance level of that test. Under the null hypothesis,

- C_+ and C_- follow a binomial distribution with parameters N and $1/2$.
- S_- and S_+ follow approximately a binomial distribution with parameters N and $\alpha/2$.
- $S_?$ follows approximately a binomial distribution with parameters N and $1 - a$.

Breakpoint Calculation

A breakpoint in the relationship between time and another variable (e.g., mean polysemy) was computed using the R package *strucchange* [42, 41] assuming that the curves contain a single breakpoint (setting the parameter *breaks* to 1). This R package implements the algorithm described by [1] for simultaneous estimation of multiple breakpoints, but in our case we are just looking for one breakpoint. The breakpoint is the value of the predictor (children's age or time

in our case) where the error is minimized in a double linear regression fit. In all cases, we get breakpoints with an error of ± 0.1 months according to a confidence level of 95%.

Acknowledgements

This research work has been supported by the SGR2014-890 (MACDA) project of the Generalitat de Catalunya, and MINECO project APCOM (TIN2014-57226-P).

References

[1] Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

[2] Jaume Baixeries, B. Elvevåg, and Ramon Ferrer-i-Cancho. The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE*, 8(3):e53227, 2013.

[3] D. Bickerton. *Language and species*. Chicago University Press, 1990.

[4] R. Brown. *A first language: the early stages*. Harvard University Press, Cambridge, MA, 1973.

[5] W. J. Conover. *Practical nonparametric statistics*. John Wiley & Sons, Inc., New York, 1999. 3rd edition.

[6] E. M. Cross and W. W. Chaffin. Use of the binomial theorem in interpreting results of multiple tests of significance. *Educational and Psychological Measurement*, 42:25–34, 1982.

[7] Scott Crossley, Tom Salsbury, and Danielle McNamara. Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59(2):307–334, 2009.

[8] Scott Crossley, Tom Salsbury, and Danielle McNamara. The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3):573–605, 2010.

[9] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: properties and pitfalls. In M. A. H. Dempster, editor, *Risk management: value at risk and beyond*, pages 176–223. Cambridge University Press, Cambridge, 2002.

[10] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Pres, Cambridge, MA, 1998.

[11] Ramon Ferrer-i-Cancho. The optimality of attaching unlinked labels to unlinked meanings. *Glottometrics*, 36:in press, 2016.

[12] Dedre Gentner. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In Stan A. Kuczaj II, editor, *Language development: Vol. 2. Language, thought and culture*, chapter 11, pages 301–334. Lawrence Erlbaum Associates, Hillsdale, 1982.

[13] Dedre Gentner. *Why verbs are hard to learn*, pages 544–564. Action meets word: How children learn verbs. Oxford University Press, 2006.

[14] J. D. Gibbons and S. Chakraborti. *Nonparametric statistical inference*. Chapman and Hall/CRC, Boca Raton, FL, 2010. 5th edition.

[15] Judith C. Goodman, Philip S. Dale, and Ping Li. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, 35(3):515–531, 2008.

[16] J. Harris, R. M. Golinkoff, and K. Hirsh-Pasek. Lessons from the crib for the classroom: how children really learn vocabulary. In S. B. Neuman and D. K. Dickinson, editors, *Handbook of early literacy research*, volume 3, pages 49–65. Guilford Press, NY, 2011.

[17] Antoni Hernández-Fernández, Bernardino Casas, Ramon Ferrer i Cancho, and Jaume Baixeries. Testing the robustness of laws of polysemy and brevity versus frequency. In P. Král and C. Martín-Vide, editors, *4th International Conference on Statistical Language and Speech Processing (SLSP 2016). Lecture Notes in Computer Science 9918*, pages 19–29, 2016.

[18] Thomas T. Hills, Josita Maouene, Brian Riordon, and Linda B Smith. The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, 63:259–273, 2010.

[19] Thomas T. Hills, Mounir Maouene, Josita Maouene, Adam Sheya, and Linda Smith. Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, 20:729–739, 2009.

[20] Ramon Ferrer i Cancho. Optimization models of natural communication. *http://arxiv.org/abs/1412.2486*, 2015.

[21] Ramon Ferrer i Cancho and Antoni Hernández-Fernández. The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4(1):45–55, 2013.

[22] R. Jackendoff. Possible stages in the evolution of the language capacity. *Trends in Cognitive Science*, 3(7):272–279, 1999.

[23] Patricia K. Kuhl. Brain mechanisms in early language acquisition. *Neuron*, 67(5):713–727, 2010.

[24] B. MacWhinney. *The CHILDES project: tools for analyzing talk*, volume 2: the database. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition, 2000.

[25] P. Matychuk. The role of child-directed speech in language acquisition: a case study. *Language Sciences*, 27:301–379, 2005.

[26] C. McDonough, L. Song, K. Hirsh-Pasek, R. M. Golinkoff, and R. Lannon. An image is worth a thousand words: why nouns tend to dominate verbs in early word learning. *Developmental Science*, 14:181–189, 2011.

[27] C. B. Mervis. Child-basic object categories and early lexical development. In U. Neisser, editor, *Concepts and conceptual development: echological and intellectual factors in categorization*, pages 201–233. CUP, New York, 1987.

[28] George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[29] Brigitte Nerlich, Zazie Todd, Vimala Hermann, and David D. Clarke, editors. *Polysemy: Flexible Patterns of Meaning in Mind and Language*. Mouton de Gruyter, Berlin, Germany, 2003.

[30] Steven Pinker. *The Language Instinct*. William Morrow and Co., New York, NY, 1994.

[31] Brandon C. Roy, Michael C. Frank, Philip DeCampa, Matthew Millera, and Deb Roy. Predicting the Birth of a Spoken Word. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41):12663–12668, 2015.

[32] Mattew Saxton. *Child Language: Acquisition and Development*. Sage publications, 2010.

[33] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*, 1994.

[34] C. E. Snow. Mothers' speech to children learning language. *Child Development*, 43(2):549–566, 1972.

[35] Mahesh Srinivasan and Hugh Rabagliati. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152, 2015.

[36] M. K. Stamer and M. S. Vitevitch. Phonological similarity influences word learning in adults learning Spanish as a foreigh language. *Bilingualism and Cognition*, 15:490–502, 2012.

[37] H. L. Storkel. Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25:201–221, 2004.

[38] H. L. Storkel, J. Armbruster, and T. P. Hogan. Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, 49:1175–1192, 2006.

[39] Michael Tomasello. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4(4):156–163, 2000.

[40] George Yule. *The Study of Language*. Cambridge University Press, Cambrigde, UK, 3rd edition, 2006.

[41] Achim Zeileis, Christian Kleiber, Walter Krämer, and Kurt Hornik. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1-2):109–123, 2003. Special Issue in Honour of Stan Azen: a Birthday Celebration.

[42] Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002.

[43] George K. Zipf. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33:251–256, 1945.

[44] George K. Zipf. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge (MA), USA, 1949.